JOURNAL OF MULTIVARIATE ANALYSIS 16, 21–53 (1985)

A Comparison of Kriging with Nonparametric Regression Methods*

S. J. YAKOWITZ AND F. SZIDAROVSZKY

University of Arizona

Communicated by M. Rosenblatt

"Kriging" is the name of a parametric regression method used by hydrologists and mining engineers, among others. Features of the kriging approach are that it also provides an error estimate and that it can conveniently be employed also to estimate the integral of the regression function. In the present work, the kriging method is described and some of its statistical characteristics are explored. Also, some extensions of the nonparametric regression approach are made so that it too displays the kriging features. In particular, a "data driven" estimator of the expected square error is derived. Theoretical and computational comparisons of the kriging and nonparametric regressors are offered. © 1985 Academic Press, Inc.

1. BACKGROUND AND SCOPE OF THIS STUDY

Specialists in hydrology, mining, petroleum engineering, and other geoscience-based subjects have recently exhibited considerable interest and enthusiasm for a methodology known as "kriging." To name only a few recent (mostly water-resource oriented) works, we mention in this regard, Bakr *et al.* [2], Chirlin and Dagan [37], David [6], Delhomme [7, 8], Dendrou and Houstis [9], Hughes and Lettenmaier [20], Gambolati and Giampiero [16], Gambolati and Volpi [15], Gelhar *et al.* [18], Huijbregts [21], Journel [23, 24], Journel and Huijbregts [22], and Villeneuve *et al.* [45]. Alldredge and Alldredge [1] offers a bibliography of mining-oriented kriging references, numbering in the hundreds. The name "Kriging" derives, according to Journel [23] from Krige [25], where the basic idea was first outlined. Matheron [31] should be credited with its early dissemination.

In the present section, we will set forth the statistical problems which the kriging method is intended to solve, and in Section 2, we will reveal the

Received November 15, 1982; revised April 28, 1983.
AMS 1980 subject classification: primary 62H15, 62J05.
Key Words and Phrases: Nonparametric regression, geostatistical methods.

* Some of the results of this work have been announced in Yakowitz and Szidarovszky [52].

0047-259X/85 $3.00

popular kriging algorithms themselves and derive their properties. It turns out that there are certain unsatisfactory theoretical aspects of kriging, and yet prior to the present study, they appear to be the only methods appropriate for the problems in their domain. However, we intend to show that methods of nonparametric regression are certainly relevant and competitive.

In Section 3, we have provided an extension of the kernel nonparametric regression approach to enhance its applicability to kriging problems. Section 4 constitutes a brief summary of the theoretical developments of this study. In particular, it offers a comparison of the salient properties of kriging with the corresponding properties of the nonparametric regression approach. In the concluding section (Section 5), results of some computer studies are described.

Let $f(x)$ and $n(x)$ be real-valued functions defined on a domain χ in R^d. Suppose $\{(x_i, y_i)\}_{i=1}^N$ is a sequence of "noisy" function pairs; that is, suppose

$$y_i = f(x_i) + n(x_i), \qquad 1 \leqslant i \leqslant N. \tag{1.1}$$

The interpretation is that $f(x)$ is a function whose values are to be estimated and $n(x)$, a "noise," represents a random function, the distribution of which is unrelated to $f(x)$. We discuss two problems which are central to the kriging literature.

Let $x^* \in \chi$ be specified. It may or may not be among the sample pairs. On the basis of the sample pairs $\{(x_i, y_i)\}_{i=1}^N$,

Problem (a)—provide an estimate $f_N(x^*)$ of $f(x^*)$, and

Problem (b)—provide an estimate of the expected squared error

$$E[(f_N(x^*) - f(x^*))^2 | x_1, ..., x_N]. \tag{1.2}$$

Remarks. The goal of problem (a) coincides with the objectives of "nonparametric regression" methods, but to our knowledge, investigators in this latter area have not concerned themselves with problem (b). Because practitioners desire to estimate piezometric head in oil and water aquifers or the grade of an ore body as a function of position, the dimension d of the domain χ is often 2 or 3.

One sometimes wishes to estimate functionals of $f(x)$, particularly integrals of $f(x)$. For through integrals, one can estimate the total weight of metal to be extracted from an ore body occupying a given region, on the basis of imperfect assay estimates of the grade at distinct locations. Therefore, at various junctures in this study, we will remark on generalizations of kriging and nonparametric regression to the functional case.

Problems (a) and (b) seem to have their roots in the forestry and geostatistics literature. In fact, it seems that "geostatistics" is almost synomymous with kriging. (It is to be noted that the recent books by Henley [19] and Ripley [36] exhibit a broader perspective, however.) We have no doubt that these problems are important and interesting. In this connection, in a review of a geostatistics book, Watson [46] has written, "The time is certainly ripe for a more serious attack on the estimation of the earth's resources,...."

2. Introduction to the Kriging Methodology and Convergence Analysis

In the kriging approach, it is presumed that $f(x)$ and $h(x)$ are stochastic processes uncorrelated from one another. More specifically, it is assumed that $f(x)$ is an *intrinsic random function* (IRF); that is, for some functions $\{\phi_i(x)\}_{i=1}^{J}$ known to the user and constants $a_1,..., a_J$, for all x, h such that x, $x+h \in \chi$,

$$E[f(x+h)-f(x)] = \sum_{j=1}^{J} a_j(\phi_j(x+h)-\phi_j(x)) \tag{2.1}$$

and, independently of x, with "var" signifying "variance,"

$$\tfrac{1}{2}\operatorname{var}[f(x+h)-f(x)] = \gamma(h). \tag{2.2}$$

If $J=1$ and $\phi_1(x)=1$, then $f(\cdot)$ is a stochastic process with stationary increments. In the more general case that $J>1$, one says that the intrinsic random function has "drift." The drift functions are often taken to be monomials up to some specified degree.

The constants $\{a_j;\ 1 \leqslant j \leqslant J\}$ and the function $\gamma(h)$ are unknown quantities. In what follows, it is presumed always that $J \leqslant N$. The function $\gamma(h)$ is called the *variogram*. Even in the case in which the mean $E[f(x)]$ exists and is known to be constant in x (i.e., $J=1$, $\phi_1(x)=1$), the hypothesis of "intrinsic random function" is weaker than second-order stationarity. For example, Brownian motion is an intrinsic random function, but it is well known to be a nonstationary process. We note that in other contexts, stochastic process structure has been superimposed on function spaces to give a framework for derivation of inference rules (e.g., Kushner [26] for sequential seach for a function maximum, and Ferguson [12], Leonard [27], and Whittle [48] by different routes, in nonparmetric Bayesian density estimation).

The kriging method is composed of two activities, (i) inferring the variogram from the data and (ii) assuming that the inferred variogram is

indeed exact, providing a best linear unbiased estimator and associated error variance.

Activity (ii) is a standard least-squares problem, and is consequently by far the best understood of the two facets of kriging. There are some inconsistencies in the fundamental definitions and results in the kriging literature. For example, the definitions of "intrinsic random function" given by David [6] and Matheron [30] do not coincide. The term "nugget effect" in the kriging literature refers to some mechanism for accounting either for measurement noise or extensive local variation in the function $f(\cdot)$ itself, such variation representing, for example, the inhomogeneity of an ore sample. This multiple use has not been carefully distinguished by kriging authors, and there has been resulting discrepancy in the mathematical representations. The equations for kriging in the presence of noise as given by Rendu [34], for example, agrees with our calculations, but differs from formulas offered by other authors (e.g., Journal [23]). In view of these inconsistencies, we have elected to derive the "universal kriging" equations for prediction with known variogram from first principles.

2.1. *Linear Estimation with Known Variograms*

To begin with, suppose the noise term in (1.1) is zero. Let us assume that the variogram $\gamma(h)$ and the mean function components $\{\phi_i(x)\}$, of the expectation (2.1) are given. The assumption that one of these functions, say ϕ_1, is 1, seems to be a standard and perhaps unavoidable assumption in view of the intrinsic random function hypothesis. To begin with, let us discuss the solution of Problem (a). The objective is to choose the parameters $\{\lambda_i\}_{i=1}^N$ on the basis of the data $\{(x_i, y_i)\}_{i=1}^N$, so that the linear estimator

$$f_N(x^*) = \lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_N y_N \tag{2.3}$$

minimizes

$$E[(f(x^*) - f_N(x^*))^2], \tag{2.4a}$$

subject to the constraint that

$$E[f_N(x^*)] = E[f(x^*)]. \tag{2.4b}$$

In view of the assumed form (2.1) of the mean value function, a necessary and sufficient condition for the unbiasedness equation (2.4b) to hold, regardless of the drift coefficients a_j, $1 \leqslant j \leqslant J$, is that

$$\sum_{i=1}^{N} \lambda_i \phi_j(x_i) = \phi_j(x^*), \qquad 1 \leqslant j \leqslant J. \tag{2.5}$$

Equation (2.5) with $\phi_1 = 1$, implies that

$$\sum_{i=1}^{N} \lambda_i = 1. \tag{2.6}$$

Use this fact, with (2.4b), to conclude that, with "cov" signifying "convariance" and $f^* = f(x^*)$,

$$\begin{aligned} E\left[\left(f^* - \sum_{i=1}^{N} \lambda_i y_i\right)^2\right] &= \operatorname{var}\left(f^* - \sum_{i=1}^{N} \lambda_i y_i\right) \\ &= \operatorname{var}\left(\sum \lambda_i (f^* - y_i)\right) \\ &= \sum_i \sum_j \lambda_i \lambda_j \operatorname{cov}[(f^* - y_i), (f^* - y_j)] \\ & \qquad (1 \leqslant i, j \leqslant N). \end{aligned} \tag{2.7}$$

Now observe that

$$\begin{aligned} \operatorname{cov}[(f^* - y_i), (f^* - y_j)] &= \tfrac{1}{2}[-\operatorname{var}((f^* - y_i) - (f^* - y_j)) \\ &\quad + \operatorname{var}(f^* - y_i) + \operatorname{var}(f^* - y_j)] \\ &= -\gamma(x_i - x_j) + \gamma(x^* - x_i) + \gamma(x^* - x_j). \end{aligned} \tag{2.8}$$

One makes these substitutions into (2.7) and after some algebra, sees that the Lagrange multiplier technique applied to the quadratic programming problem of minimizing $E[(f(x^*) - f_N(x^*))^2]$ subject to (2.5) yields

$$\sum_{k=1}^{N} \lambda_k \gamma(x_i - x_k) \gamma(x_i - x^*) + \sum_{j=1}^{J} \mu_j \phi_j(x_i), \quad 1 \leqslant i \leqslant N, \tag{2.9a}$$

$$\sum_{i=1}^{N} \lambda_i \phi_j(x_i) = \phi_j(x^*), \qquad 1 \leqslant j \leqslant J. \tag{2.9b}$$

The variables μ_j are the Lagrange multipliers. Journel [23] calls the above linear equation the *universal kriging system.* Alternative formulations of quadratic programming problem solutions found in Chapter 10 of Fletcher [14] may be of use in the context of kriging. Note that the solution may fail to exist because (2.9b) cannot be satisfied.

From substitution according to (2.9) into (2.7), one concludes that the mean-squared prediction error is given by

$$E[(f^* - f_N(x^*))^2] = \sum_{i=1}^{N} \lambda_i \gamma(x^* - x_i) - \sum_{j=1}^{J} \mu_j \phi_j(x^*). \tag{2.10}$$

If the noise term $n(x)$ in (1.1) is present and has zero mean, one accounts for its presence by noting that, because it is presumed uncorrelated from the f-process,

$$\begin{aligned}\operatorname{cov}((f^*-y_i),(f^*-y_j)) &= \operatorname{cov}((f^*-f_i-n_i),(f^*-f_j-n_j)) \\ &= \operatorname{cov}((f^*-f_i),(f^*-f_j)) + \operatorname{cov}(n_i,n_j).\end{aligned}$$

In the above equation, we have, of course, intended that f_i and n_i signify $f(x_i)$ and $n(x_i)$. As a result of the above, one readily sees that in the presence of noise (2.9a) should be replaced by (2.9′a):

$$\sum_{k=1}^{N} \lambda_k(\gamma(x_i-x_k)-\operatorname{cov}(n(x_i),n(x_k))) = \gamma(x_i-x^*) + \sum_{j=1}^{J} \mu_j\phi_j(x_i), \qquad 1\leqslant i\leqslant N. \tag{2.9′a}$$

With coefficients γ_i and μ_j obtained from (2.9a), formula (2.10) still gives the expected squared estimation error.

Let us generalize the universal kriging equations to the task of estimation of functionals of the intrinsic random function. For simplicity, assume the noise is 0. Let $L(\cdot)$ denote a measureable linear functional on the universe of $f(\cdot)$ satisfying the following conditions:

(i) If $1(x)$ is the function equal to 1 on χ, then $L(1(x))=1$,

(ii) L commutes with the expectation operator, and

(iii) $\operatorname{var}(L(f)-f(x))$ and $L(\phi_j)$ exist for every point x in χ, and for $1\leqslant j\leqslant J$.

If $f(\cdot)$ is continuous in the mean, then L can represent normalized Riemann integrals over bounded domains B. That is, $L(f)=D^{-1}\int_B f(x)\,dx$, where $D=\int_B dx$. This is the customary application. The condition (i) is an annoying (because it precludes differentiation, for example) consequence of the intrinsic random function assumption. If one assumes only that $f(\cdot)$ itself is a stationary process and uses covariances instead of variograms, then the domain of application can be widened.

Toward deriving the universal kriging equations for estimation of the functional $L(f)$, proceed as in the point estimation case, writing $G_N=\sum_{i=1}^{N}\lambda_i y_i$ as the linear estimator. The problem is, of course, to choose the λ_i's to find the unbiased minimizer of

$$\begin{aligned}E[(L(f)-G_N)^2] &= \operatorname{var}[(L(f)-G_N)^2] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\operatorname{cov}((L(f)-y_i),(L(f)-y_j)).\end{aligned}$$

Now confirm that

$$\operatorname{cov}((L(f)-y_i),(L(f)-y_j)) = \tfrac{1}{2}[-\operatorname{var}((L(f)-y_i)-(L(f)-y_j)) + \operatorname{var}(L(f)-y_i)+\operatorname{var}(L(f)-y_j)].$$

The first variance term on the right in the above is $-\gamma(x_i - x_j)$, and after some effort, one confirms that

$$\tfrac{1}{2}(\operatorname{var}(L(f)-y_i)+\operatorname{var}(L(f)-y_j)) = L_x L_y \gamma(x-y) + L\gamma(x-x_i) + L\gamma(x-x_j), \tag{2.11}$$

where $L_x L_y$ indicates that the operation L is to be taken with respect to both variables x and y. With these representations in terms of the variogram, it is at last a simple matter to see that the Lagrange condition for minimizing $E[(L(f)-G_N)^2]$, subject to $E[G_N] = E[L(f)]$ is

$$\begin{aligned} \sum_{k=1}^{N} \lambda_k \gamma(x_i - x_k) &= L(\gamma(x-x_i)) + \sum_{i=1}^{J} \mu_j \phi_j(x_i), \qquad 1 \leqslant i \leqslant N, \\ \sum_{i=1}^{N} \lambda_i \phi_j(x_i) &= L\phi_j, \qquad 1 \leqslant j \leqslant J, \end{aligned} \tag{2.12}$$

and the expected squared error of the constrained least squares estimate is

$$E[(L(f)-G_N)^2] = \sum_{i=1}^{N} \lambda_i L(\gamma(x-x_i)) - \left(\sum \mu_j L\phi_j + L_x L_y(\gamma(x-y))\right). \tag{2.13}$$

Equations (2.9) and (2.10) are specializations of (2.12) and (2.13) to L determined by $L(f) = f(x^*)$.

2.2. *Inference of the Variogram*

The task of inferring a covariance function or power spectral density from data is known by experienced statisticians to be somewhat delicate, and one which furthermore requires a considerable quantity of data. The subtleties of the covariance inference problem translate directly to the task of inferring a variogram from data.

There are some very real difficulties with variogram estimation in the published kriging applications. To avoid effects of "nonstationarity," practitioners tend to have a single variogram apply only to a relatively small region χ of domain points of $f(x)$. Moreover, they have not developed procedures to ascertain whether the intrinsic random function hypothesis is tenable for their applications. A particular difficulty is that in the bounded

TABLE I

A Listing of Popular Variogram Families

Generalized Linear	$\gamma_\theta(h) = \omega \|h\|^a, \quad 0 < a < 2$	
Spherical	$\gamma_\theta(h) = \omega\left[\frac{3}{2}\frac{\|h\|}{a} - \frac{1}{2}\left(\frac{\|h\|}{a}\right)^3\right]$	$\|h\| < a$
	$= \omega,$	$\|h\| > a$
Exponential	$\gamma_\theta(h) = \omega[1 - \exp(\|h\|/a)]$	
Gaussian	$\gamma_\theta(h) = \omega[1 - \exp(\|h\|^2/a^2)]$	

Note. $\theta = (a, \omega)$ is the parameter in each case.

domain case, ergodic theorems are inapplicable to the task of demonstrating consistency. More will be said on this point in Section 2.4.

We now concern ourselves with outlining the present practice with regard to variogram inference. The customary procedure is to choose a parametric family of variograms from the five or six popular families mentioned in the literature, and then to select the variogram from the chosen family which agrees best, in some sense, with the data $\{(x_i, y_i)\}_{i=1}^N$. We list in Table I some of the prominent variogram families, with $\|h\|$ signifing Euclidean norm:

One should note that not just any family of functions will do for parametric variograms. Specifically, in view of (2.8), the function

$$Q(x, y) = \gamma(x - x_0) - \gamma(y - x_0) - \gamma(y - x)$$

must be nonnegative definite on χ^2, for any x_0 in χ, in order to insure that $\text{var}(\sum a_j(f(x_0) - f(x_j))$ be nonnegative regardless of choice of a_j's and x_j's. As in Table I, practitioners prefer "isotropic" variograms, i.e., variograms which depend only on the Euclidean norm of h. If the associated process $f(\cdot)$ happens to be stationary, then from discussion in Wong [50, Chap. 7, Sect. 3] there are very stringent conditions that the variogram must satisfy. These conditions appear difficult to check.

There seems to be no consensus in the literature on methodology for the selection of a parametric family from Table I on the basis of an observed sample $\{(x_i, y_i)\}_{i=1}^N$. Some heuristic approaches for inferring the variogram of $f(x) + n(x)$ are proposed by David [6]. This differs from the variogram of $f(x)$ by a constant term in the white noise case. Concerning the task of selection of the member $\gamma_\theta(h)$ the foremost criteria seem to be (i) least

squares, (ii) cross validation, and (iii) a geometric procedure (David [6]). In the least squares approach, one selects the parameter θ^* so as to minimize

$$I_1(\theta) = \sum_v (\gamma_N(h_v) - \gamma_\theta(h_v))^2,$$

the index v running over some finite collection of arguments h_v and $\gamma_N(h)$ being some sample approximation to the variogram, such as

$$\gamma_N(h) = (2N(h))^{-1} \sum_{j=1}^{N(h)} (y_j - y_{j(h)})^2,$$

where $j(h)$ is an index selected so that $x_j - x_{j(h)} \simeq h$ and $N(h)$ is the number of such points selected. The cross-validation approach to parameter selection is as follows. Let $P(\theta, x_j)$ be the universal kriging estimate of $f(x_j)$ on the basis of the sample points $\{(x_i, y_i)\}_{i \neq j}$ and parametric variogram $\gamma_\theta(h)$. One then chooses θ^* to minimize the squared error of the predicted values, which is

$$I_2(\theta) = \sum_{j=1}^{N} (y_j - P(\theta, x_j))^2.$$

If "drift" is thought to be present (i.e., if $\phi_j, j > 1$, in (2.1) is not zero), these approaches entail some serious conceptual difficulties. Matheron [30, Chap. 4] has addressed these difficulties.

Practitioners insist, quite rightly, that one should not select a variogram entirely algorithmically, but that attention should be paid to past experience with similar geostatistical data.

2.3. *Estimation Convergence with Correct Variogram*

With the exception of studies by Matheron [29, 30], the literature of kriging tends to be practical and pragmatic. Major issues of consistency and convergence rates have not been investigated. In the developments to follow, we attempt to obtain initial results in these areas.

Let us begin our analysis of convergence of the kriging estimate under the simplest of conditions by assuming that

(i) The observations are noiseless ($n(x_i) = 0$).

(ii) $\gamma(0) = 0$, and γ is continuous in a neighborhood of the origin.

(iii) There is no "drift"; that is, $J = 1$ and $\phi_1 = 1$.

(iv) The "true" variogram is known.

THEOREM 2.1. *Let χ be the domain of the intrinic random function $f(x)$*

and assume that the conditions above are in force. If the infinite sequence $\{x_i\}$ *is dense at* $x^* \in \chi$ *and for* $f_N(x^*)$ *determined by* (2.9),

$$E[(f(x^*) - f_N(x^*))^2] \to 0 \qquad \text{as} \quad N \to \infty. \tag{2.14}$$

Proof. In view of assumption (iii), for every i, $y_i = f(x_i)$ is itself an unbiased linear estimator of $f(x^*)$, and so for $N \geqslant 1$.

$$E[(f(x^*) - f_N(x_i))^2] \leqslant E[(f(x^*) - f(x_i))^2] = 2\gamma(x^* - x_i).$$

Let $x^*(N)$ denote the member of $\{x_i\}_{i=1}^N$ which is closest to x^*. By the assumption that x^* is an accumulation point of $\{x_i\}$, $x^*(N) \to x^*$ as $N \to \infty$, and therefore

$$\begin{aligned} E[(f(x^*) - f_N(x^*))^2] &\leqslant E[(f(x^*) - f(x^*(N))^2] \\ &= 2\gamma(x^* - x^*(N)). \end{aligned} \tag{2.15}$$

The proposition follows by observing that, in light of property (ii), $\gamma(x^* - x^*(N))$ must converge to 0. The bound given by (2.15) may be of some practical interest in itself. The idea in this proof will serve us in other developments. ∎

The Brownian motion process affords an example of a situation in which the best estimate is not consistent unless x^* is an accumulation point of the sample points $\{x_i\}$. For Brownian motion is Markov, and the best estimate of $f(x^*)$ will depend only on the points $(x_a, f(x_a))$ and $(x_b, f(x_b))$, where x_a is the largest domain sample less than x^* and x_b the smallest sample greater than x^*.

In some applications, the following assertion may be cogent.

COROLLARY. *Assume that the hypotheses of Theorem* 2.1 *are in force and additionally that* χ *is open and has finite Lebesgue measure* $\gamma(h)$ *has a continuous second derivative, and the samples* $\{x_i\}$ *are identically and independently distributed on* χ *with* pdf *bounded away from* 0 *in a neighborhood of* x^*. *Then*

$$E[(f(x^*) - f_N(x^*))^2] = O(N^{2/d}), \tag{2.16}$$

d being the dimension of the space containing χ, *expectation here being with respect to* $f(\cdot)$ *and* $\{x_i\}$.

Proof. Since $\gamma(h)$ is an even function, the gradient at $h = 0$ must be 0, and we have

$$\begin{aligned} \gamma(x^* - x^*(N)) = (\tfrac{1}{2})(x^* - x(N))^T \gamma^{(2)}(O)(x^* - x(N)) \\ + o(\|x^* - x^*(N)\|^2). \end{aligned} \tag{2.17}$$

It is known (e.g., Yakowitz *et al.* [53, p. 1299]) that under the independent, uniformly distributed sample case, for all points $x^* \in \chi$ and some constant C_1,

$$E[\|x^* - x^*(N)\|^2] = O(N^{2/d}), \qquad N = 1, 2, ... \tag{2.18}$$

From the argument in that reference, one can conclude that (2.18) holds whenever the pdf is bounded away from 0 in a neighborhood of x^*. The Corollary now follows from (2.17) and (2.18). ∎

On pp. 50–51 of Ripley [36] there is some discussion about the "smoothness" of the estimated surface and the differentiability of the sample function itself. These results do not impinge on our consistency findings above inasmuch as Ripley assumes that the number of observations N is fixed. We warn the reader that Ripley's claim that the estimated surface is at least as smooth as the sample function is not generally true. (The sample functions of the random process $f(x) = \sin(x + \theta)$, θ uniformly distributed on $[0, 2\pi]$, are infinitely differentiable, and yet the covariance function of this process has $\alpha = 1$, in Ripley's terminology.)

From our experience in groundwater analysis, where the domain points correspond to well locations, the hypotheses of the corollary are of some use. On the other hand, for some ore sampling strategies, it may be more reasonable to assume that the x_i's form a grid of similar-sized rectangles. For such regular patterns, one may conclude that (2.18) is true without expectations.

We will now discuss convergence of the kriging estimate when accounting for drift. Assume that x^* is a limit point of $\{x_i\}$. Assume furthermore that for some subsequence $x_{n_1}, ..., x_{n_J}$ the matrix $\Phi \triangleq \{\phi_i(x_{n_j})\}_{i,j=1}^J$ is nonsingular. (Otherwise, there is no hope of being able to obtain estimates satisfying (2.5) for arbitrary $\phi_i(x^*)$ values.) Let us further assume that the drift functions have continuous first derivatives. For $N > n_j$, define the linear estimate

$$\tilde{f}_N(x^*) = (1 - \alpha_N) f(x^*(N)) + \alpha_N \sum_{i=1}^{J} \lambda_N^i f(x_{n_i}), \tag{2.19}$$

where $x^*(N)$ is, as before, the nearest neighbor (among the first N samples) to x^*, and

$$\alpha_N \triangleq \|x^* - x^*(N)\|. \tag{2.20}$$

In order to assure that the constraint condition (2.5) holds, we set $\boldsymbol{\phi}(x) = (\phi_1(x), ..., \phi_J(x))^T$ and determine $\boldsymbol{\lambda}_N = (\lambda_N^1, ..., \lambda_N^J)^T$ by

$$\alpha_N \Phi \boldsymbol{\lambda}_N = \boldsymbol{\phi}(x^*) - (1 - \alpha_N)\, \boldsymbol{\phi}(x^*(N)). \tag{2.21}$$

The consistency of the estimate $\tilde{f}_N(x^*)$ will follow from the argument leading to (2.15) if only we can show the sequence $\{\boldsymbol{\lambda}_N\}$ remains in a bounded region. Toward that end, note that after taking a Taylor's series expansion of $\boldsymbol{\phi}(x^*) - \boldsymbol{\phi}(x^*(N))$ and dividing α_N, we may rewrite (2.21) as

$$\boldsymbol{\Phi}\boldsymbol{\lambda}_N = \boldsymbol{\phi}(x^*(N)) - (1/\alpha_N)\,\nabla\boldsymbol{\phi}(x^*(N) - x^*) + 1/\alpha_N o(\|x^* - x^*(N)\|), \quad (2.22)$$

where the matrix

$$\nabla\boldsymbol{\phi} \triangleq \begin{pmatrix} \nabla\phi_1(x^*) \\ \vdots \\ \nabla\phi_J(x^*) \end{pmatrix}. \quad (2.23)$$

From (2.22), we see that $\boldsymbol{\lambda}_N$ remains bounded when α_N is chosen according to (2.20). In fact,

$$\|\boldsymbol{\lambda}_N\| \leqslant \|\boldsymbol{\Phi}^{-1}\|[\|\nabla\boldsymbol{\phi}\| + \|\boldsymbol{\phi}(x^*)\|] + O(1). \quad \blacksquare$$

If the conditions of the corollary to Theorem 2.1 are in force, the convergence rate of that corollary apply in the drift case also since the convergence of $x^*(N)$ to x^* is not influenced by the constraints.

Our attention now turns to the case that noise $n(x_i)$ is present in the observation law (1.1). For simplicity, assume that $J = 1$ and $\phi_1 = 1$. If $n(\cdot)$ is a continuous function, then apparently consistent identification of $f(x^*)$ is not possible since local samples cannot distinguish between the effects of signal and noise. However, the linear estimate provided by the universal kriging equations is an appropriate procedure and in fact coincides with what is known to communication engineers as a "smoothing filter." If $\{n(x_i)\}$ are independent variables, then, as we now demonstrate, under some circumstances, consistent estimation of $f(x^*)$ is possible. Toward verifying this assertion, as in earlier arguments, we find a linear estimator whose properties are understood and then appeal to the fact that since the kriging estimate is optimal in the least squares sense, it must be at least as good as the estimator under consideration.

For the particular task at hand of verifying consistency in the presence of independent noise, it is sufficient to call attention to the fact that Stone [43] has discussed a general class of nonparametric regression rules,

$$\hat{f}_N(x^*) = \sum_{i=1}^{N} y_i w_{i,N}(x^*; x_1, \ldots, x_N).$$

The weights $w_{i,N}$ can be taken to add to 1 (i.e., $\sum_{i=1}^{N} w_{i,N} = 1$), so the unbiasedness condition (2.5), with $J = 1$ holds. His results imply that if x^*

and x_i are obtained from i.i.d. observations, and if sample functions $f(x)$ of $f(\cdot)$ are measurable, and provided that weight functions $w_{i,N}(\cdot)$ satisfy certain natural properties, then $\hat{f}_N(x^*) \to f(x^*)$ in the mean.

Toward applying Stone's results to the issue of consistency of kriging estimates in the noisy observation case, let $\bar{f}(\cdot)$ denote a realization of the intrinsic random function $f(\cdot)$. Then if $y_i = \bar{f}(x_i) + n(x_i)$, the sequence $\{(x_i, y_i)\}$ constitutes i.i.d. observations and the hypotheses of Stone's convergence result holds, provided a few technical assumptions of little practical concern hold. Since kriging gives the least squares unbiased estimator, we conclude that

$$E[(f_N(x^*) - \bar{f}(x^*))^2] \to 0, \qquad \text{as} \quad N \to \infty.$$

Here, expectation is with respect to the $\{x_i\}$ and $\{n(x_i)\}$ sequences. It may be concluded that if the noise measurements and the sample functions $\bar{f}$ are uniformly bounded, then convergence occurs without the condition of sample function $\bar{f}$; alternatively, without the boundedness assumption, one can assert that convergence in the mean is assured outside an f-set of arbitrarily small positive measure. From results in Section 3, it may be seen that if one is willing to assume that the sample functions are twice-continuously differentiable, then convergence in the mean is on the entire f-space without the set qualification.

For certain specific NPR estimates, rates of convergence are known (e.g., Fisher and Yakowitz [13], Parthasarathy and Bhattacharya [33], Sacks and Spiegelman [38], Schuster and Yakowitz [40], Mack and Silverman [28]). The strongest results related to convergence of point NPR estimates known to us are that of Schuster [39[for one-dimensional x_i's, and for d-dimensional x_i's, the result to be demonstrated in the next section, that for $f_N(x^*)$ the kernel NPR estimate for $\bar{f}(x^*)$, that with some provisos to be specified,

$$E[(f_N(x^*) - \bar{f}(x^*))^2] = O(n^{-(1/(d/4+1))}). \tag{2.24}$$

In evaluating the convergence statements concerning kriging up to this point, it should be emphasized that they are valid only if $f(\cdot)$ really is an intrinsic random function and the variogram and drift functions are *known perfectly*.

Let us examine the issue of convergence of estimates of linear functionals $L(f(\cdot))$ of the intrinsic random function $f(\cdot)$. The assumptions (i) through (iii) introduced earlier in connection with the universal kriging equation (2.12) for linear functionals will be presumed in force. The implications of the preceding results on convergence of point estimators are fairly evident once we establish part (b) of the following statement.

THEOREM 2.2. *Let $f_N(x)$ denote the universal kriging estimate of $f(x)$, on the basis of data $\{(x_j, y_j)\}_{i=1}^N$, with $\sigma_N(x)$ giving the standard deviation of this estimate. Then*

$$\text{(a)} \qquad G_N = L(f_N(\cdot)) \tag{2.25}$$

is the kriging estimate for $L(f)$, and

$$\text{(b)} \qquad E[(L(f) - G_N)^2] \leqslant L(\sigma_N(\cdot))^2. \tag{2.26}$$

Proof. The idea for proving (a) is to show that G_N as determined by (a) satisfies the Lagrange conditions. Toward this end, let $\lambda(x)$ and $\mu(x)$ denote the vectors of coefficients λ_i, $1 \leqslant i \leqslant n$, and μ_j, $1 \leqslant j \leqslant J$, of the solution of (2.9a) and (2.9b) for $x^* = x$. From examination of these equations, one concludes that they bear the representation

$$\begin{pmatrix} \boldsymbol{\lambda}(x) \\ \boldsymbol{\mu}(x) \end{pmatrix} = K(\mathbf{w}(x)), \tag{2.27}$$

where K is a matrix which does not depend on x, and the vector $\mathbf{w}(x)$ is determined in an evident manner by $\{\gamma(x - x_i)\}$ and $\{\phi_j(x_j)\}$. From the representation (2.27), it is clear that the solution vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ of the linear functional kriging equation (2.12) are given by

$$\begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{pmatrix} = KL(w) = L \begin{pmatrix} \boldsymbol{\lambda}(x) \\ \boldsymbol{\mu}(x) \end{pmatrix}. \tag{2.28}$$

That is, $\boldsymbol{\lambda} = L(\boldsymbol{\lambda}(x))$, whence if $y' = (y_1, \ldots, y_N)$.

$$G_N = y' L(\boldsymbol{\lambda}(x)) = L(f_N(\cdot)).$$

To demonstrate part (b), as before, use the subscript to distinguish the function that L is acting on and note that

$$\begin{aligned} &E[(L(f - f_N))^2] \\ &\quad = E[(L_x(f - f_N)(x))(L_y(f - f_N))(y)] \\ &\quad = L_x L_y E[(f - f_N)(x)(f - f_N)(y)] \leqslant L_x L_y \sigma_N(x)\, \sigma_N(y) \\ &\quad = (L(\sigma_N(\cdot))^2. \quad \blacksquare \end{aligned} \tag{2.29}$$

We have difficulty interpreting the discussion at the bottom of p. 51 of Ripley [36], but it may be that these remarks foresee part (a) and its proof, in the case that L represents integration.

2.4. *Estimation Convergence with Incorrect Variograms*

For simplicity, assume the no drift case. It is farily clear that if the variogram is in error, there is little hope of estimating $E[(f(x^*)-f_N(x^*))^2]$ correctly.

As has been noted earlier, even in the noiseless case, there is typically no consistent variogram estimator based on $\{(x_i,\bar{f}(x_i))\}_{i=1}^n$ for x_i in a bounded domain χ and $\bar{f}$ a fixed sample of an intrinsic random function f. In short, we are unaware of consistent variogram estimates, even if $\gamma(h)$ is known to be a member of a given family such as listed in Table I. The essential trouble is that for consistent variogram estimation, the observed sample function $\bar{f}(x)$, $x\in\chi$, must somehow contain decisive information about $\operatorname{var}[f(x)-f(x+h)]$. To cite a contrived example where this certainly is not the case, suppose that χ is the unit interval and $f(x)$ is a birth–death process with birth intensity = death intensity = λ. Then λ must somehow be estimated by some finite number of "steps" in $\bar{f}(x)$. But the problem of inferring λ can be translated into the problem of inferring an exponential parameter from a finite random sample, and this cannot be done perfectly. The most promising case might be $f(x)$ an ARMA process of known order on χ. But even here, to our knowledge, there are no techniques in the identification literature which purport arbitrarily accurate parameter estimation from a finite sample segment. The measurement error and computational stability difficulties would appear to be enormous. We note that Doob [11, p. 531] has written, "It is important to note the $R(t)$ [the covariance function]..., in general cannot be determined from a knowledge of sample functions in a finite interval."

On the other hand, as we will soon demonstrate, under certain circumstances, the kriging estimate will converge, with an increasing number of samples, to the correct value, even when the variogram is not correct. An interpretation of these remarks is that the kriging method is in some cases effective for estimating values on the basis of noisy samples, but that the associated error estimate need not be consistent. This interpretation is borne out by our simulation studies. The fact that the estimate of the squared error need not become more accurate with increasing data is significant because kriging practitioners and their clients place great value on the error estimation feature.

EXAMPLE. In this example, we show that it is possible for the kriging predictor to be exact, while the variogram (and hence the error estimate) may contain significant error. Suppose $\gamma_2 = b\gamma_1$, where b is any positive constant. If $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_N)$ is the minimizer of (2.7), subject to the constraints (2.5), with $\gamma=\gamma_1$, then $\boldsymbol{\lambda}$ will also be the constrained minimizer of (2.7) with $\gamma=\gamma_2$. Thus if a presumed variogram is to much as approximately propor-

tional to the correct one, the estimate $f_N(x^*)$ will be reliable. But from (2.10), one sees that (ignoring the drift term) the error estimate under γ_2 will differ from that under γ_1 by the scale factor b.

Let $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{\lambda}^{(2)}$ be the solutions of the universal kriging equation (2.9a, b) under variograms γ_1 and γ_2, respectively. Suppose that for some positive number δ and all h, $|\gamma_1(h) - \gamma_2(h)| < \delta$. From a standard numerical analysis formula (e.g., Szidarovszky and Yakowitz [44, p. 214], we have that for $\delta < \|A\| \backslash \Gamma(A)$,

$$\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\| < \Gamma(A)\|\boldsymbol{\lambda}^{(1)}\| \, \delta/(\|A\| - \delta\Gamma(A)), \tag{2.30}$$

where A is the matrix determined in connection with (2.26) and

$$\Gamma(A) = \|A\| \, \|A^{-1}\|$$

is the condition number. Some insight into the potential perniciousness of variogram error can be inferred from (2.30) by considering that the linear equation associated with least squares problems frequently is ill-conditioned because of collinearity effects. This phenomenon is evidenced by large condition number $\Gamma(A)$.

Our objective now is to show that under some circumstances, the kriging point estimate will be consistent even when the inferred variogram is not a good approximation of the true variogram. Our results are not as comprehensive as we would like, but they would appear to have some potential value as a guide for choosing a variogram family when physical imperatives are lacking. The idea guiding the arguments to follow is that many predictors can be demonstrated to asymptotically place increasing proportions of their prediction weights close to the prediction point x^*, as the domain points $\{x_i\}$ become dense at x^*. An unfortunate restriction of the analysis to follow is that we must assume that $f(\cdot)$ itself is a stationary second order random function. For now, consider the noiseless case, and let $R(h)$ be the covariance function. That is,

$$R(h) = \operatorname{cov}[f(x+h), f(x)].$$

Assume that $R(\cdot)$ has a spectral density $S(w)$. For any square-integrable functions $g(x)$ and $h(x)$ of d variables, we will let (g, h) denote the inner product over R^d defined by

$$(g, h) = \int g(x) \, h(x) \, dx,$$

and the least-squares predictor f_N of $f(x^*)$, determined by $R(\cdot)$ and the observed pairs $\{(x_j, y_j)\}_{i=1}^N$ can therefore be represented as (f, λ_N), where

$$\lambda_N(x) = \sum_{i=1}^{N} \lambda_i \delta(x - x_i),$$

$\delta(x - x')$ being the Dirac delta function at x'.

LEMMA 2.1. *If $\{x_i\}$ is dense at x^* and for some positive numbers q and C, $\liminf |w|^q S(w) \geqslant C$ as $|w| \to \infty$, then for any continuous function $g(x)$ of bounded support as $N \to \infty$,*

$$\lim(\lambda_N, g) = g(0). \tag{2.31}$$

Proof. We may as well take $x^* = 0$. Letting tildes denote d-dimensional Fourier transforms of the indicated functions, and noting that $\tilde{\delta}(x) = 1$, a standard spectral representation formula (e.g., Rozanov [37], Yagiom [51], or for $d > 1$, Wong [50]) asserts

$$E[(f(0) - f_N)^2] = \int |1 - \hat{\lambda}_N|^2 S(w)\, dw. \tag{2.32}$$

Zemanian [55, Chap. 4] defines testing functions of rapid descent to be infinitely differentiable functions satisfying the condition that for any positive M, $\phi(w)$ (and its derivatives) satisfy $\|w\|^M |\phi(w)| \leqslant C(M)$, all w, $C(M)$ depending only on ϕ and M. One concludes that for some positive number K, $KS(w) > |\tilde{\phi}(w)|$, every w. From this fact, Eq. (2.32), and the fact (Theorem 2.1) that $E[(f(0) - f_N)^2] \to 0$, we can conclude that for any testing function ϕ of rapid descent,

$$\lim(1 - \hat{\lambda}_N, \phi) = 0 \qquad \text{as} \quad N \to \infty.$$

Thus, according to terminology of Zemanian, $\hat{\lambda}_N \to 1$, over testing functions of rapid descent. But there is a Fourier transform continuity theorem (Zemanian [55, p. 187]) that assures us that as a consequence, $\lambda_N \to \delta(x)$, over testing functions of rapid descent. Finally, it is known that continuous functions $g(x)$ with bounded support can be uniformly approximated arbitrarily closely by testing functions. Therefore (2.31) follows. ∎

Now we view the lemma from a kriging standpoint.

THEOREM 2.3. *Let χ_1 be a bounded subset of χ containing x^* and $\{x_j\}$, of which x^* is presumed to be a limit point. Assume f_N is the least-squares predictor of $f(x^*)$ determined by $\{(x_i, y_i), 1 \leqslant i \leqslant N\}$ and $R(\cdot)$, which is presumed to satisfy the hypotheses of the lemma. Finally, suppose $f(\cdot)$ is a*

stochastic process which is continuous, a.s. *Then* $f_N \to f(x^*)$, a.s., *regardless of the covariance structure governing the process* $f(\cdot)$.

Proof. Let $\bar{f}(\cdot)$ be a sample function, and $g(\cdot)$ any continuous function with bounded support that agrees with $\bar{f}(\cdot)$ on χ_1. Then by the lemma as $N \to \infty$, ∎

$$(\lambda_N, g) = g(x^*) = \bar{f}(x^*). \tag{2.33}$$

The preceding analysis did not attach constraints to $\{\lambda_i^N\}$. The reader will be able to confirm that convergence as in (2.33) is still assurred if the weights are constrained to add to 1. More generally, the important case of covariance functions $R(\cdot)$ having rational spectral densities can be seen to satisfy the convergence property of Theorem 2.3. Toward this end, note that for the conclusion of the lemma to hold, it is sufficient that for any $e > 0$, there exist a positive constant $C(e)$ and a set $A(e)$ such that

$$\|w\|^q S(w) > C(e), \qquad w \notin A(e)$$

and $\int_{A(e)} dw < e$. For under these circumstances, letting $\phi_{\max}$ denote the maximum of $|\phi(w)|$, we have

$$\lim_N |(1 - \hat{\lambda}_N, \phi)| < \phi_{\max} e \qquad \text{as} \quad N \to \infty,$$

and the result follows inasmuch as the choice of e is arbitrary. Such spectral densities then satisfy the lemma hypothesis.

The Gaussian family is a somewhat amusing case. It does not satisfy the lemma hypothesis; but furthermore it does not satisfy the usual criterion of "regular" (e.g., Doob [11]) second-order process. That is, for the spectral density determined by the Gaussian variogram,

$$\int_{-\infty}^{\infty} [\ln(S(w))/(1 + w^2)]\, dw = -\infty.$$

For such processes, $f(x^*)$ can be exactly determined (by Taylor's expansions) from f-values on any interval, no matter how far away from x^*. Wiener [49] has argued that only regular processes make sense as physical models.

We currently have no results for the case in which white noise is present. It is not enough to know that the "λ-weight" asymptotically concentrates close to x^*. One must also be assured that the estimate does some sort of averaging. In particular, we have mentioned that if $d = 1$, the exponential variogram implies a predictor that applies all weight to only two points. This would not be a sensible method in the noisy case. A technical difficulty is that in the white noise case, $R(\cdot)$ is discontinuous at 0, and yet essentially all spectral analysis theory assumes continuity at 0 (e.g., Doob [11, p. 518]).

In the above estimation convergence analysis, it was assumed that the same fixed, possibly incorrect, covariance function $R(\cdot)$ was used at each stage N in computation of the predictor f_N of $f(x^*)$. In practice, one would "update" the estimated convariance function as more data became available, thus obtaining a sequence $\{R_N(\cdot)\}$ of "random" covariance functions. Unfortunately, such a procedure seems to greatly complicate the convergence analysis. In particular, there seems to be no assurance that $E_N[(f_N - f(x^*))^2] \to 0$ as $N \to \infty$, the expectation E being with respect to the points $\{x_i, 1 \leqslant i \leqslant N\}$ and the assumed covariance R_N. While the preceding convergence results can be extended to the "random" covariance function case under certain stringent assumptions, at present we regard the situation as a (perhaps unfillable) lacuna in kriging theory.

3. Application of Nonparametric Regression

The purpose of the present section is to reveal extensions of nonparametric regression theory which makes this approach more suited to Problem (b), Section 1. The particular nonparametric regression (NPR) method to be investigated here is the kernel estimator proposed by Watson [47]. The two developments revealed here are (i) a formula for the asymptotic expected squared error and (ii) a data-based approximation of the mean squared error. The discussion closes by showing that the asymptotic convergence of the NPR estimates is, in a certain sense, optimal.

Comparing the hypotheses of nonparametric regression to those of kriging, there are three salient distinctions:

(1) The "target function" $f(\cdot)$ (in this section termed "regression function") is not presumed to be stochastic, but rather some arbitrary but unknown function which we will presume to be twice continuously differentiable.

(2) The points $\{x_i\}$ in (1.1) are presumed to be observations of an i.i.d. random variable X.

(3) The noise components $n(x_i)$ of (1.1) are presumed independent and to have finite variance here, but in contrast to the preceding section, we will allow that the variance may depend on position.

As a result of these assumptions, one may view the data $\{(x_i, y_i)$: $1 \leqslant i \leqslant N\}$ as being i.i.d. observations of a random pair $(X, f(X) + N(X))$, X being a d-dimensional random vector, $f(\cdot)$ a function known to be twice continuously differentiable, and $N(\cdot)$ a random variable with variance $v(x)$ continuously indexed by the argument. These assumptions are by no means universal in the nonparametric regression literature. In the section to follow, other avenues will be cited.

The kernel estimator constructed on the basis of a random sample $\{(x_i, y_i)\}_{i=1}^N$ is

$$f_N(x) = \left[\sum_{i=1}^{N} y_i\, k((x_i - x)/a_N)\right] / D_N(x), \tag{3.1}$$

where $D_N(x) = \sum_{i=1}^N k((x_i - x)/a_N)$, a_N is a positive number, and $k(\cdot)$ is a probability density function chosen by the statistician.

Since in kriging, squared-error is the essence, our analysis at this point is directed toward establishing the behavior of $E[(f_N(x) - f(x))^2]$ as the number N of observations increases. Toward that end, let $h(x, y)$ and $g(x)$ be the pdf's of (X, Y) and X, respectively. Let $w(x) = \int y\, h(x, y)\, dy$, thus $f(x) = w(x)/g(x)$, and define for some d-tuple x^* and $1 \leqslant i \leqslant N$,

$$U_{Ni} = k((x^* - X_i)/a_N)/a_N^d, \tag{3.2}$$

$$V_{Ni} = Y_i U_{Ni}, \tag{3.3}$$

$$U_N = 1/N \sum U_{Ni} \colon 1 \leqslant i \leqslant N, \tag{3.4}$$

$$V_N = 1/N \sum V_{Ni} \colon 1 \leqslant i \leqslant N. \tag{3.5}$$

Throughout this section, we will assume that the kernel pdf $k(u)$ is selected so as to satisfy the properties (i) to (iv) below:

(i) $k(u)$ and $\|u\|\, k(u)$ are bounded,

(ii) $\int uk(u)\, du = 0$,

(iii) $\int \|u\|^2 k(u)\, du < \infty$,

(iv) the functions $g(x)$ and $w(x)$ are twice continuously differentiable and the second partial derivatives of $g(x)$ are bounded,

(v) the second moment of Y is finite.

The pdf of the multivariate normal law satisfies properties (i)–(iii).

The convergence facts we will need are given in the statement below. The hypothesis and conclusion of this theorem are similar but not coincident with results announced (but not proven) in Collomb [4].

THEOREM 3.1. *Let d be the dimension of the sample vectors $x_1, x_2, \ldots$, and assume $g(x^*) > 0$. Then*

(a) $\operatorname{var}(V_N)$ *and* $\operatorname{var}(U_N)$ *are both* $O(1/(N a_N^d))$,

(b) $(E[V_N] - w(x^*))$ *and* $(E[U_N] - g(x^*))$ *are both* $O(a_N^2)$. (3.6)

(c) *If* $a_N = N^{-(1/(d+4))}$; *then*

$$E[(f_N(x^*) - f(x^*))^2] = O(N^{-(1+d/4)^{-1}}).$$

Proof. This proof is very much inspired by developments of Schuster [39]. Thus part (a) is essentially formula (1) in the proof of his Lemma 1, but extended here to d variables. In particular, after a change of variables to $u = (x_i - x^*)/a_N$ we have

$$E[(U_{Ni})^2] = a_N^{-d} \int k(-u)^2 \, g(x^* - a_N u) \, du$$

$$= (g(x^*)/a_N^d) \left[\int k^2(u) \, du + O(a_N) \right].$$

Similarly, one may confirm that

$$E[U_{Ni}] = g(x^*) \left[\int k(u) \, du + O(a_N) \right].$$

Now use that the variables are uncorrelated to get $\text{var}(U_N) = O((Na_N^d)^{-1})$. The demonstration for $\text{var}(V_N)$ proceeds in the same fashion. The proof of part (b) is essentially that of the first part of Lemma 2 in Schuster [39]. Thus after the change in variable, and use of assumed property (ii) above,

$$E[U_{Ni}] - g(x^*) = \int k(-u)[\, g(x^* - a_N u) - g(x^*)] \, du$$

$$\leqslant (a_N^2/2) \sup_x \| g''(x) \| \int \|u\|^2 k(u) \, du = O(a_N^2).$$

Clearly, U_N and U_{Ni} have the same expectation. The analysis of $E[V_N]$ proceeds in a similar fashion.

Toward demonstration of (c), define E_N to be the event that

$$U_N > (\tfrac{1}{2}) \, g(x^*) \qquad \text{and} \qquad |V_N| < 2 \, |w(x^*)|.$$

From strong consistency results of Silverman [41] and Mack and Silverman [28], the complement of E_N can occur but finitely many times. Now under E_N,

$$|f_N(x^*) - f(x^*)| = |(V_N \, g(x^*) - U_N w(x^*))/U_N \, g(x^*)|$$

$$\leqslant (2/g(x^*)^2)(|\, w(x^*)(U_N - g(x^*))|$$

$$+ g(x^*)|\, V_N - w(x^*)|).$$

Part (c) now is easily seen to be a consequence of (a) and (b). ∎

Our attention now turns to derivation of a data-based estimate of the mean squared error of the NPR point estimate $f_N(x)$, i.e.,

$$E[(f_N(x)-f(x))^2 | X_j = x_j,\ 1 \leqslant j \leqslant N]. \tag{3.7}$$

Observe that since the terms $\{V_{Ni}\}_i$ in (3.5) are uncorrelated,

$$\operatorname{var}(f_N | \{x_i\}_{i=1}^N) = \sigma_N^2(x) \triangleq (1/D_N(x)^2) \times \left(\sum_{i=1}^N E[n(x_i)^2]\, k^2((x-x_i)/a_N) \right). \tag{3.8}$$

The only term in (3.8) which is not known to the statistician is $E[n(x_i)^2]$. But this can be approximated from the sample by defining α to be any positive number less than 1 and defining

$$\hat{E}[n(x_i)^2] = 1/N^\alpha \left(\sum_{j \in S(i,N)} (y_j - f_N(x_j))^2 \right), \tag{3.9}$$

where $S(i, N)$ is the set of indices of the N^α nearest neighbors in $\{x_k\}_{k=1}^N$ of x_i. Since in view of (3.6), $f_N(x)$ is converging in N to $f(x) = E[Y|X = x]$, and since with probability 1, the radii of the sets $S(j, N)$ become vanishingly small as $N \to \infty$, it is evident that the estimate

$$\hat{\sigma}_N^2(x) \triangleq (1/D_N(x))^2 \sum_{i=1}^N \hat{E}[n(x_i)^2]\, k^2((x-x_i)/a_N), \tag{3.10}$$

satisfies the relation

$$\hat{\sigma}_N^2(x)/\sigma_N^2(x) \to 1 \qquad \text{as} \quad N \to \infty, \text{ i.p.} \tag{3.11}$$

Note that the estimator $\hat{\sigma}_N^2(x)$ depends solely on the statisticians choices of $k(\cdot)$ and $\{a_N\}$, and the observed sequence $\{(x_j, y_j)\}$. From the theorem, one may conclude that if a_N tends to zero slightly faster than $(1/N)^{(1/(d+4))}$, then the variance error part of (a) will dominate, yet need not seriously degrade the rate of convergence in (3.6). Under this circumstance, $\hat{\sigma}_N^2(x)$ will be an asymptotically accurate (in the sense of 3.11)) estimate of the expected square error (3.7).

Of course, asymptotic results must be further examined to provide guidance in the moderate sample size case. A procedure to be discussed in Section 5 involves using cross-validation to provide guidance in the selection of the bandwidth parameter a_N. Presumably, in cross-validation, the data has helped us in finding empirically the bandwidth for which the rms error due to estimator variance approximates the bias error. A pragmatic procedure might be to assume that the trade-off has been reached and then simply

double the estimated variance error $\hat{\sigma}_N^2(x)$ to get an estimate for the squared error of prediction. Alternatively, before using (3.10) one could decrease the parameter a_N a little bit from the cross-validation "optimal" value to (hopefully) force the variance term to dominate. Our findings in controlled simulation experiments such as to be reported in Section 5 are that when the measurement error $n(X_i)$ is not insignificant, the error estimate $\hat{\sigma}_N^2(x)$ does give a good idea of the actual error.

One can confirm that for any $\delta > 0$, as $a_N \to 0$, the contribution in (3.10) of terms x_i such that $\|x - x_i\| > \delta$, becomes negligible, and in practice, we have found that

$$\hat{\sigma}_N^2(x) = \hat{E}[n(x)^2](1/D_N(x))^2 \left(\sum_{i=1}^{N} k^2((x - x_i)/a_N) \right) \tag{3.12}$$

gives a reliable approximation of the error variance. Similarly, one can show that for any points x^1, x,

$$\hat{\rho}_N(x, x^1) = \left[(1/D_N(x)\, D_N(x^1)) \sum_{i=1}^{N} k((x - x_i)/a_N)\, k((x^1 - x_i)/a_N) \right] \times [\hat{E}(n(x)^2)\, \hat{E}(n(x^1)^2)]^{1/2} \tag{3.13}$$

is an asymptotically accurate estimate of $\operatorname{cov}(m_N(x), m_N(x^1))$. This relationship is useful in applying the NPR approach to linear functionals of the regression function, as is now discussed.

Let us briefly turn our attention to NPR estimation of linear functionals $L(f)$ of the "target function" $f(x)$. In contrast to conventions in kriging, NPR methodology does not call for unbiased estimation (although our kernel estimator (3.3) does satisfy the condition (2.6) that $f_N(x)$ be a convex combination of the y_i's). Thus the stipulation required in the kriging approach that $L(1) = 1$ need not apply here, and L can be presumed to be a differential operator, for example.

The natural approach is to choose the "kernel" density $k(\cdot)$ so that $L(k)$ is well defined, and then set

$$L(f_N) = \sum_{i=1}^{N} y_i L(k(x - x_i)/a_N)/D_N(x)). \tag{3.14}$$

Under this circumstance, letting the subscript of L denote the variable on which the operation is applied, we have

$$\begin{aligned} E[(Lf - Lf_N)^2] &= E[(L_x(f(x) - f_N(x))(L_y(f(y) - f_N(y))] \\ &= L_x L_y \operatorname{cov}(f(x) - f_N(x), f(y) - f_N(y)) \\ &\cong L_x L_y \hat{\rho}_N(x, y) \leqslant (L(\hat{\sigma}_N))^2. \end{aligned} \tag{3.15}$$

The final consideration regarding the kernel estimator concerns a certain optimality property. In view of (3.6) and the Chebyshev inequality, one can conclude that for $r^* = 2/(d+4)$, and for any regression function $m(x)$ and noise process $n(x)$ satisfying the theorem hypothesis, if $a_N \to 0$ proportionally to $N^{-(d+4)^{-1}}$, then for $r = r^*$,

$$\lim_{C\to\infty} \limsup_{N} P[|m_N(x^*) - m(x^*)| > C/N^r] \to 0. \tag{3.16}$$

Thus, in the terminology of Stone [42], the NPR estimate achieves convergence rate r^*. But according to the theorem of that work, for any NPR estimator of a twice continuously differentiable regression function of d independent variables, $r^* = 2/(d+4)$ is the optimal rate: there is no estimator for which (3.16) holds for some $r > r^*$ over all such regression functions $f(x)$.

4. A Comparison of Convergence Properties of Kriging and Nonparametric Regression

Assume that the intrinsic random function (IRF) hypothesis holds, and there is no drift ($J = 1$, $\phi_1 = 1$). If the domain points are chosen randomly and if, with probability 1, the sample functions $\bar{f}(x)$ of the IRF are continuous then the NPR estimate $f_N(x)$ converges to $\bar{f}(x)$ in the mean (Devroye and Wagner [10]). If the sample IRF's are twice continuously differentiable, with probability 1, then Theorem 3.1 gives convergence rates. Cramer and Leadbetter [5] give conditions on covariance functions under which versions of the stochastic process can be assured to have various smoothness properties.

Toward addressing Problem (b) of Section 1, we have provided error formulas (3.10), (3.12), and (3.15) which are asymptotically accurate provided only that the sample functions are continuous at x^*. These statements have been predicated on the assumption that the $\{x_i\}_{i=1}^N$ values are actually a random sample. However, under fairly lenient assumptions, Schuster and Yakowitz [40, Theorem 2] have shown in the univariate case that $f_N(x)$ converges uniformly in x to $\bar{f}(x)$ provided only that the x_i's are dense. Gasser and Müller [17] have derived sharp convergence bounds for several nonparametric regression schemes for certain classes of real nonrandom sequences $\{x_i\}$. Undoubtedly such results can be extended to bear on kriging-type problems more forcefully, and citations of related results (especially concerning the Priestly–Chao estimator) are to be found in the above references.

Now it is clear that if the variogram is known exactly, because the kriging

estimator is the best unbiased linear estimator, then the expected square error of the kriging estimator $f_N(x)$ is no greater than that of the NPR estimator, which is also linear and unbiased (in the sense that the weights add to unity). On the other hand, in the noisy case, it is not known at this point whether its asymptotic convergence rate is faster than the NPR rate given in Theorem 3.1. In summary, when the IRF hypothesis is true and the variogram is known to the statistician, the kriging estimate is at least as good in the least squares sense as the NPR estimate, and its error estimators (2.10) and (2.13) are exact, whereas the NPR error estimators are only asymptotically accurate.

In the case that the IRF hypothesis cannot be relied upon or when the assumed variogram is incorrect, in the noiseless case under fairly lenient assumptions the kriging estimate converges to $f(x^*)$ if only the assumed variogram is that of a regular stationary stochastic process. These matters were discussed at the close of Section 2. On the other hand, nothing can be said in the general case about the accuracy of the kriging estimate of error when the variogram is not correct. Moreover, if noise is present, or if the variogram assumes noise is present (by being discontinuous at 0), then the earlier convergence analysis does not apply and at this stage, we have no reason to think that the kriging estimator converges to the desired value $f(x^*)$. In Table II we have loosely summarized our findings. The reader should refer back to the relevant discussion for more precise convergence statements.

TABLE II

Summary of Convergence Results

No Noise	Kriging Correct Variogram	Incorrect Variogram	NPR
Estimator	*Y*	*Y*	*Y*
Error Estimate	*Y*	*N*	
Noise			
Estimator	*Y*	*Y*	*Y*
Error Estimate	*Y*	*N*	*Y*

Note. *Y* = Yes, convergence is established, or error estimate is correct. *N* = No, there are counterexamples to convergence.

5. A Computational Study

The authors have engaged in extensive computational experimentation with the view of comparing the performances of the kriging and kernel NPR approaches on various types of simulated data sets. In this section we report the applications of these methods to several sets which we consider to be representative of distinctly different types of modelling assumptions. To begin with, details of the implementation of the methods are revealed, and then we present the results of the experimentation.

5.1. *The Kriging Algorithm*

Our extensive experimentation with one and two-dimensional domain sets χ have included all the families listed in Table II. Our impression is that the exponential variogram performs about as well as any of the others, and for some data sets it performs much better than the worst case. An advantage to the exponential variogram model is that it is relatively easy to simulate realizations of the associated stochastic process.

In our computer studies, we allowed for the possibility of independent noise observations $n(x_i)$ in (1.1). Thus in addition to inferring the parameters a and ω of the exponential variogram, one must infer σ^2, where $\sigma^2 = E[n(x)^2]$. Our procedure was to simultaneously choose a, ω, and σ^2 to provide the least squares fit with the sample convariance function. This least-squares procedure (described in Section 2) has also been employed by David [6]. The best (in this least squares sense) variogram having thus been determined, the kriging estimates of $f(x^*)$ and the estimate of square error were found by the universal kriging system (2.9a) and the square error formula (2.10).

5.2. *The Kernel NPR Algorithm*

In the case of the kernel estimator (3.1), there are two objects left to the user's discretion, namely the "bandwidth" parameter a_N and the kernel function $k(\cdot)$, itself. The standard normal density function was chosen as the kernel; it has sufficient smoothness to satisfy the hypothesis of NPR convergence theorems. For bandwidth selection, the following cross-validation technique was used: For each number j, $1 \leqslant j \leqslant N$, define $f_{N,j}(x, a)$ to be the value of the NPR estimate (3.3) constructed from the data $\{(x_v, y_v)\}_{v \neq j}$ and "bandwidth" parameter a. Then define

$$H(a) = \sum_{j=1}^{N} (y_j - f_{n,j}(x, a))^2.$$

Finally, a uniform search is used to choose the value a^* which minimizes $H(a)$. This number a^* is used as a_n. Such cross-validation procedures in this

context have been shown by Rice [35] to have attractive properties for evenly-spaced values $\{x_i\}$.

The reader will note that by our implementations, both the kriging and NPR algorithms are completely automatic. We used exactly the same code for all the data sets to be described.

5.3. *Experiments in Univariate Case*

Whereas in practice, kriging is typically a multivariate enterprise, because of our predilection for graphical representations, we have found one-dimensional domain case studies informative for computational experimentation. The domain χ of the four experimental real "target" functions was the unit interval, the domain points $\{x_i\}$ being independently and uniformly chosen from χ, and the number N of observations was 50. The first function was an observation of the Ornstein–Uhlenbeck process. Thus,

$$d/dt f(t) = -a f(t) + W(t),$$

where $W(t)$ is white noise. The associated variogram for $f(t)$ is $\omega(1 - \exp(-a|h|))$, i.e., it is the exponential variogram. Newman and Odell [32] describe how to simulate realizations of $f(t)$ which are exact, within the constraints of machine error and to the extent that one is able to provide independent Gaussian observations. These are approximated by the Box–Muller algorithms (described in Yakowitz [54]) using the CDC random number generator RANF.

Here and in the ensuing experiments, both the kriging and NPR algorithms operate on exactly the same data sets. (We note that, technically speaking, the "target" function $f(x)$ in this example does not satisfy the conditions of our Theorem 3.1 because it is not sufficiently differentiable. On the other hand, results in Devroye and Wagner [10] assure us of convergence of $f_N(x)$ to $f(x)$ (but no rate is guaranteed) under conditions which include this example.)

Function 2 corresponds to Function 1, except additive noise $n(x_i)$ as in

TABLE III

Summary of rms Approximation Errors, 1D Case

	Krig	Kernel NPR
Function 1	0.159	0.182
Function 2	0.208	0.208
Function 3	0.219	0.107
Function 4	0.360	0.181

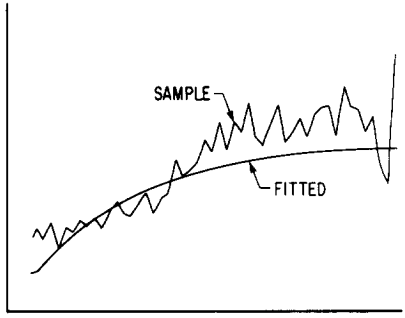

FIG. 5.1. Sample variogram and its fitted exponential approximation.

(1.1) is added to the function $f(x_i)$ in generating the data points $\{(x_i, y_i)\}_{i=1}^{50}$. Here we are still within the hypothesis for the kriging scheme. In geostatistical terminology, discussed in Section 2, we have an intrinsic random function with the "nugget" effect. The simulated noise is i.i.d. normal, with mean 0 and standard deviation of 0.5, a noise process also used for the remaining function experiments. The third target function is $f(x) = \sin(3x)$, and thus the intrinsic random function hypothesis is not active. For the fourth function, the unit step with step at $x = 0.5$ serves as target. Table III summarizes the actual rms estimation error at 50 evenly-spaced domain points z_i. That is, each entry is given by

$$\left[\frac{1}{50}\sum_{i=1}^{50}(f_N(z_i) - f(z_i))^2\right]^{1/2},$$

where $f_N(z)$ is either the kriging or kernel estimate.

In Fig. 5.1, the sample variogram associated with the third function is plotted against its fitted member from the exponential family with nugget

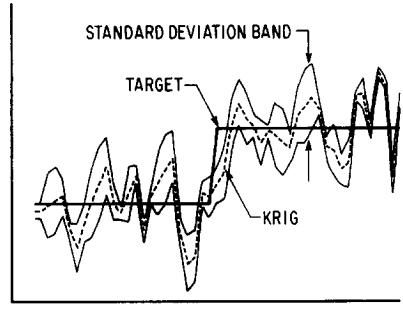

FIG. 5.2. Krige estimate of step function.

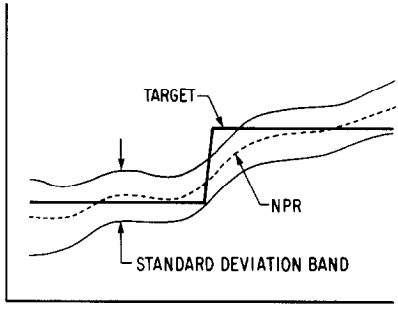

FIG. 5.3. NPR Estimate of step function.

effect accounted for. The fit appears satisfactory. Other sources of prediction error, however, are that the intrinsic random function hypothesis may fail to give adequate approximation and even if it does, the sample variogram may not be close to the theoretic variogram.

In Fig. 5.2. we have plotted the kriging estimate of the fourth function, along with its standard deviation contour against the target function. We interpret the jerkiness of the krig estimate as arising from the choice of exponential variogram family, which in the noiseless case leads to a nearest-neighbor estimate, as we discussed at the close of Section 2. Figure 5.3 gives the corresponding estimate of rms error was obtained by doubling the estimated deviation (3.12). Recall the rationale, discussed in Section 3, is that at the cross-validated bandwidth, the standard deviation ought to approximately equal the bias.

5.4. *Experiments in the Case of a Multivariate Domain*

The dimension $d = 2$, and the domain χ for the studies we now report was the unit square. As in all these cases, independent $N(0, 0.25)$ noise was added to the functional values. The domain points were chosen independently and uniformly from χ. The number of sample points N was 50, and the rms errors to be reported were obtained at the 9 evenly spaced positions $\{(i/3, j/3)\colon 1 \leqslant i, J \leqslant 3\}$. The four target functions were

$$
\begin{aligned}
f_1(x, y) &= \sin(x^* y), \\
f_2(x, y) &= \sin(2x + 2y), \\
f_3(x, y) &= 0 \qquad \text{all} \quad x, y, \\
f_4(x, y) &= 1_A(x, y)
\end{aligned}
$$

with A being the set $\{(x, y)\colon x + y > 1\}$ and 1_A being the indicator function for A. Table IV reports the rms kriging and NPR prediction errors for the data sets produced by these functions, as described above.

TABLE IV

Summary of rms Approximation Errors, 2*D* Case

	Krig	Kernel NPR
Function 1	0.234	0.105
Function 2	0.297	0.214
Function 3	0.356	0.105
Function 4	0.348	0.253

5.5. *Some Conclusions, a Disclaimer, and a Challange*

A conclusion the authors have made, and a conclusion which is exemplified if not demonstrated by the above computations, is that when the sample data does indeed satisfy the intrinsic random function hypothesis and when one does have the "true" variogram family, then the kriging estimator does perform better than the NPR approach, but only marginally better. But on the other hand, the kriging approach does not seem robust; when the data does not come from an intrinsic random function with the right variogram, the NPR approach seems consistently more reliable, especially with regard to error estimation. A side effect of our experimentation is that we were surprised at how well target functions can be approximated by a manageable number N of data points. For instance, Fig. 5.4 gives the scatter plot from which the step function was recovered in Figs. 5.2 and 5.3. Our eye did not "detect" the step function underlying the generating process.

Regarding the computational experiments and our analysis thereof, a disclaimer is in order: since kriging is an art rather than an algorithm, we cannot be assured that someone else's kriging code might not perform significantly better than ours. Commercial packages are long, expensive, presumably sophisticated, and not available to us. We would welcome the

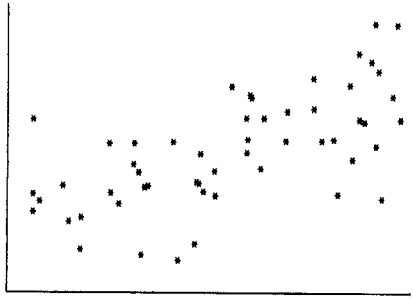

FIG. 5.4. Scatter plot of step function data.

opportunity to pit nonparametric repression techniques against one of the well-known kriging packages in a carefully designed experiment.

ACKNOWLEDGMENTS

This work is the product of evolution and labor over several years. Many kriging partisans, most notably, G. De Marsilly, J. P. Delhomme, G. Gambolati, and S. Neuman have been kind enough to patiently explain their viewpoints in conversations with the first author. Also, the first author is grateful for fruitful discussions about kriging with P. K. Bhattacharya, J. L. Denny, and E. Schuster.

This collaborative research was made possible by the NSF cooperative grant (with the Hungarian Mining Authority) ENG Int. 78-12184 and, additionally, the first author received support for this work from NSF grants ENG 76-20280, 78-07358, CME 7905010, and CEE 8110778.

REFERENCES

[1] ALLDREDGE, J. R., AND ALLDREDGE, N. G. (1978). Geostatistics, a bibliography. *Internat. Statist. Rev.*, **46**, 77–88.

[2] BAKR, A. A., GELHAR, L. W., GUTJAHR, A. L. AND MACMILLAN, J. R. (1978). Stochastic analysis of spatial variability in subsurface flows 1. Comparison of One- and Three-Dimensional Flows. *Water Resources Res.* **14** (2), 263–271.

[3] CHIRLIN, G. R. AND DAGAN, G. (1980). Theoretical head variograms for steady flow in statistical homogeneous aquifers. *Water Resources Res.* **16** (6), 1001–1015.

[4] COLLOMB, G. (1977). Quelques Propriétés de la Méthode du Noyau Por l'éstimation Non-paramétrique de la Regression en un point fixé. *C. R. Acad. Sci. Paris* **285**, 289–292.

[5] CRAMER, H. AND LEADBETTER, M. R. (1967). *Stationary and Related Stochastic Processes.* Wiley, New York.

[6] DAVID, M. (1977). *Geostatistical Ore Reserve Estimation.* Elsevier, New York.

[7] DELHOMME, J. P. (1979). Spatial variability and uncertainty in groundwater flow parameters: A geostatistical approach. *Water Resources Res.* **15** (2), 269–280.

[8] DELHOMME, J. P. (1978). Kriging in the hydrosciences. *Adv. in Water Resources*, **1** (5), 251–266.

[9] DENDROU, B. A. AND HOUSTIS, EN N. (1978). An inference-finite element model for field problems. *Appl. Math. Modelling* **2**, 109–114.

[10] DEVROYE, L. AND WAGNER, T. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8**, 231–239.

[11] DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[12] FERGUSON, T. (1973). A Bauesian analysis for some non-parametric problems. *Ann. Statist.* **1**, 209–230.

[13] FISHER, L. AND YAKOWITZ, S. (1976). Uniform convergence of the potential function algorithm. *SIAM J. Control* **14**, 95–103.

[14] FLETCHER, R. (1981). *Constrained Optimization*, Vol. II. Wiley, New York.

[15] GAMBOLATI G. AND VOLPI, G. (1979). A conceptual deterministic analysis of the Kriging technique in hydrology. *Water Resources Res.* **15** (3), 625–629.

[16] GAMBOLATI, G. AND GIAMPIERO, V. (1979). Groundwater contour mapping in Venice by stochastic interpolators. 1. Theory. *Water Resources Res.* **15** (2), 281–290.

[17] GASSER, T. AND MULLER, H. (1979). Kernel Estimation of Regression Functions. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, Eds.). Springer-Verlag, Berlin.
[18] GELHAR, W., GUTJAHR, A. L. AND NAFF, R. L. (1979). Stochastic analysis of macrodispersion in a stratified aquifer. *Water Resources Res.* **15** (6), 1387–1397.
[19] HENLEY, S. (1981). *Nonparametric Geostatistics.* Appl. Sci., London.
[20] HUGHES, J. AND LETTENMAIER, D. (1981). Data requirements for kriging: Estimation and network design. *Water Resources Res.* **17** (6), 1641–1650.
[21] HUIJBREGTS, C. J. (1975). *Regionalized Variables and Quantitative Analysis of Spatial Data.* (J. C. Davis and M. J. McCullagh, Eds.). Wiley, New York.
[22] JOURNEL, A. G. AND HUIJBREGTS, CH. J. (1978). *Mining Geostatistics.* Academic Press, New York.
[23] JOURNEL, A. G. (1977). Kriging in terms of projections. *J. Math. Geol.* **9** (6), 563–586.
[24] JOURNAL, A. G. (1974). Geostatistics for conditional simulations of ore bodies. *Econom. Geol.* **69** (5), 673–687.
[25] KRIGE, D. G. (1951). *A Statistical Approach to Some Mine Valuations and Allied Problems on the Witwaterstrand,* unpublished Master's thesis, Univ. of Witwaterstrand, South Africa.
[26] KUSHNER, H. (1964). A new method for locating the maximum point in an arbitrary multipeak curve in the presence of noise. *ASME J. Basic Engrg.* **86**, 97–106.
[27] LEONARD, T. (1978). Density estimation stochastic processes, and prior information. *J. Roy. Statist. Soc. Ser. B* **40** (2), 113–146.
[28] MACK, Y. AND SILVERMAN, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete.* **60**, 405–415.
[29] MATHERON, G. (1973). The intrinsic random functions and their applications. *Adv. in Appl. Probab.* **5**, 439–468.
[30] MATHERON, G. (1971). *The Theory of Regionalized Variables and Its Applications,* Les Cahiers du CMM. Fasc. No. 5, ENSMP, Paris.
[31] MATHERON, G. (1963). Principles of Geostatistics. *Econom. Geol.* **58**, 1246–1266.
[32] NEWMAN, T. AND ODELL, P. (1971). *The Generation of Random Variates.* Griffin, London.
[33] PARTHASARTHY, K. R., AND BHATTACHARYA, P. K. (1961). Some Limit theorems in regression theory. *Sankhyā, Ser. A* **23**, 91–102.
[34] RENDU, J. (1980). Disjunctive kriging: Comparison of theory with actual results. *Math. Geol.* **12** (4), 305–320.
[35] RICE, J. (1983). Bandwidth choice for nonparametric kernel regression. In *Proceedings of 15th Conference on the Interface of Computer Science and Statistics,* in press.
[36] RIPLEY, B. (1981). *Spatial Statistics.* Wiley, New York.
[37] ROZANOV, YU. A. (1967). *Stationary Random Processes.* Holden–Day, San Francisco.
[38] SACKS, J. AND SPIEGELMAN, C. (1980). Consistent window estimation in nonparametric regression. *Ann. Math. Statist.* **9** (2), 240–246.
[39] SCHUSTER, E. F. (1972), Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Ann. Math. Statist.* **43** (1), 84–88.
[40] SCHUSTER, E. F. AND YAKOWITZ, S. (1979). Contributions to the theory of nonparametric regressions, with applications to system identification. *Ann. Statist.* **7** (1), 139–149.
[41] SILVERMAN, B. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6**, 177–184.
[42] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** (6), 1348–1360.
[43] STONE, C. J. (1977). Consistent nonparametric regression. *Ann Statist.* **5** 595–620.

[44] SZIDAROVSZKY, F., AND YAKOWITZ, S. (1978). *Principles and Procedures of Numerical Analysis.* Plenum, New York.

[45] VILLENEUVE, J. P., MORIN, G., BOBEE, B., LEBANC, D., AND DELHOMME, J. P. (1979). Kriging in the design of streamflow sampling networks. *Water Resources Res.* **15** (6), 1833–1840.

[46] WATSON, G. S. (1977). Review of advanced geostatistics in the mining industry. *J. Amer. Statist. Assoc.* **72**, 687–688.

[47] WATSON, G. S. (1964). Smooth regression analysis. *Sanknyā Ser. A* **26**, 359–372.

[48] WHITTLE, P. (1958). On the smoothing of probability density function. *J. Roy. Statist. Soc. Ser. B.* **20**, 334–343.

[49] WIENER, N. (1964). *Time Series.* MIT Press, Cambridge, Mass.

[50] WONG, E. (1971). *Stochastic Processes in Information and Dynamical Systems.* Wiley, New York.

[51] YAGIOM, A. M. (1962). *Theory of Stationary Random Functions.* Prentice–Hall, Englewood Cliffs, N.J.

[52] YAKOWITZ, S. AND SZIDAROVSZKY, F. (1982). Regression Methods for Spatial Data. In *Proceedings of the NASA Workshop on Density Estimation and Function Smoothing.* pp. 343–385, Mathematics Department, Texas A & M University.

[53] YAKOWITZ, S., KRIMMEL, J. AND SZIDAROVSZKY, F. (1978). Weighted Monte Carlo integration. *SIAM J. on Numer. Anal.* **15** (6), 1289–1300.

[54] YAKOWITZ, S. (1977). *Computations Probability and Simulation.* Addison–Wesley, Reading, Mass.

[55] ZEMANIAN, A. H. (1965). *Distribution Theory and Transform Analysis.* McGraw–Hill, New York.